CS 294-8 Computational Biology for Computer Scientists **Spring 2003**

Lecture 5: February 4

Lecturer: Gene Myers *Scribe: Antonio Garcia*

5.1 Today's lecture

Day 5: Sequencing error characteristics
DNA assembly problem formulation
Lander-Waterman sampling analysis
Overlap/Layout/Consensus paradigm
$\rightarrow$ Sequence Comparison 101

5.2 Sequencing error characteristics

5.2.1 Errors arising from Sanger method

In whole genome shotgun sequencing, the steps were roughly as follows: take a genome sequence G, fragment it into a (roughly) uniform and random assortment of subsegments, do paired-end sequencing via the Sanger method, and feed the resulting data into some sort of software algorithm to hopefully make sense of it all. The last step is the subject of today's lecture.

With fragmented DNA separating through our capillary tube, and detected by a laser, the difference between, for example, a 10 bp segment and a 11 bp segment is a good 10%. However, between segments 500 and 501 bp, the difference is much more subtle and the two might be confused in the detection step of the capillary tube electrophoresis. The (erroneous) conclusion would then be that a given nucleotide is at the wrong position in the sequenced genome.

In addition, the incorporation location of the chain-terminating dideoxynucleotide in the Sanger reaction follows a binomial distribution, and hence is much less likely to terminate the complementary DNA copy towards the long end of the chain than at some earlier point. As a result, the peaks at the high base-pair end of the Sanger ladder spectrum are relatively faint and subject to misinterpretation.

Also, the longer (e.g. ˜500 bp) segments take longer to travel the electropheresis gel (recall the 'snake in the grass' metaphor), the peaks tend to flatten out and widen considerably, with respect to peaks from shorter segments which tend to be more compact.

Towards 800-900 bp sequences, one sees "pure mush", no matter how good the sequencing hardware. The observed error rate, on average, is anywhere from 1-3%, reaching as high as 10% in the long end of the Sanger ladder.

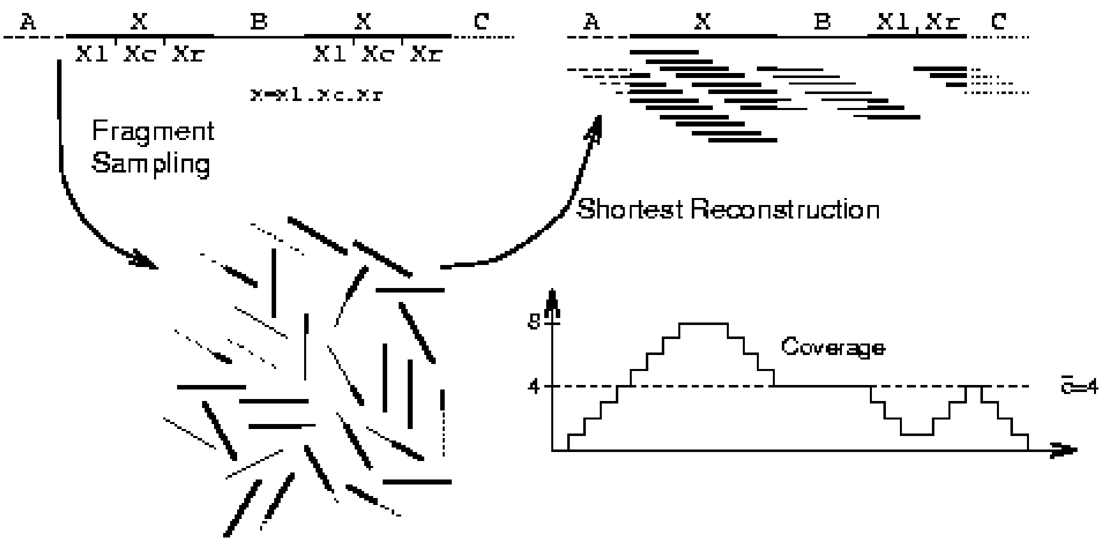

Figure 5.1: Sequencing steps: 1) a DNA sequence with segments A, B, C, and X is fragmented and then pieced together.The plots in the upper and lower right refer respectively to the resulting alignment (note skewed output sequence) and plot of the overall sequence coverage by read segments.

5.2.2 Handling error: the quality value

Due to varying error rate, an attempt was made to quantify the probable accuracy of a given position's identity. Analyzing the waveform thickness of the Sanger ladder plot and amplitude spacing of steps, we assign a corresponding number (somewhat arbitrary). The resulting metric ranges in value from 0-40, and is defined as follows:

$$Quality\,value = -10\,log_{10}(probability\,of\,error) \tag{5.1}$$

So, for example,

percent error		quality value
10%	$\rightarrow$	10
1%	$\rightarrow$	20
.1%	$\rightarrow$	30
.01%	$\rightarrow$	40

Every base pair hence carries a figure of confidence for comparison against base pairs at other positions. Saves hassle of looking at the Sanger data itself (a "chromatogram"), and lets decision-making be done numerically.

Question to Prof. Myers: Why can't you just redo sequencing and take a consensus as final answer? **Answer:** The data is over-sampled already. For example, using typical numbers of $R(reads) = 1000$, $L(length) = 500$, $G(genomelength) = 100,000$, our coverage factor is $c = \frac{RL}{G} = 8$. At Celera, declared a segment to be in the "clear range" if error≤ 1

Question to Prof. Myers: Is some of the error due to the DNA polymerase? **Answer:** No. Unless the sequence is over 1000 bp, the polymerase error is neglible.

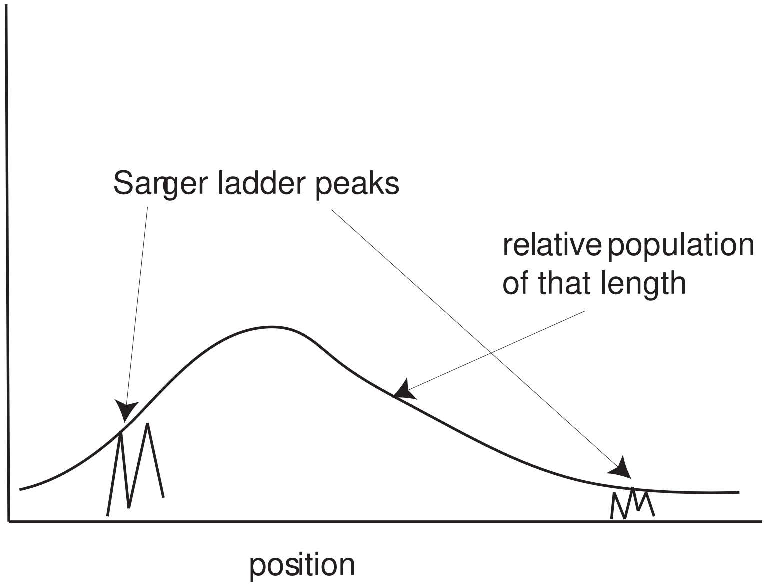

Figure 5.2: A Sanger plot

An additional source of error is the non-uniform expression of different segments of DNA during the bacteria cloning step. Some segments, for mysterious reasons concerning the bacteria's DNA and possible interference with it, do not get cloned as well as others, and hence are over (or under) represented. Likewise, at the fragmentation step, not all cuts in the sequence are necessarily equally biased. A preference for fragmentation would skew the distribution of positions within the segments.

Lastly, there is also laboratory handling error. In the paired-end sequencing step, forward and backward sequencing reactions are actually run in different plates, hence if matching is lost between complements of the same strand, the entire sequencing could be compromised. For example, at Celera, sequencing involved a 12-step process, which resulted in a 7% of human error. Even the use of barcodes on plates still involve a .5% error rate. Another example: a dud sample fails to clone in the bacteria amplification step. However, some pipeting robot's arm is contaminated, spreading successful DNA segments to an empty train, resulting in its amplification instead. At Celera, error was managed down to about 1%.

To summarize, sources of error include:

- sequencing error, due to over-representation of shorter strands, difficulty in distinguishing one basepair differences in long strands, or errors in detection in capillary separation

- cloning bias, due to interaction between DNA segment and bacteria DNA

- non-randomness in fragmentation

- mere handling error

5.3 The Lander-Waterman Equations

Consider a sequencing problem with a genome of length $G = 100,000bp$, with R reads of length L, and a coverage factor c define by $c = RL/G$.

Question 1: What percent of target genome G has been sampled?

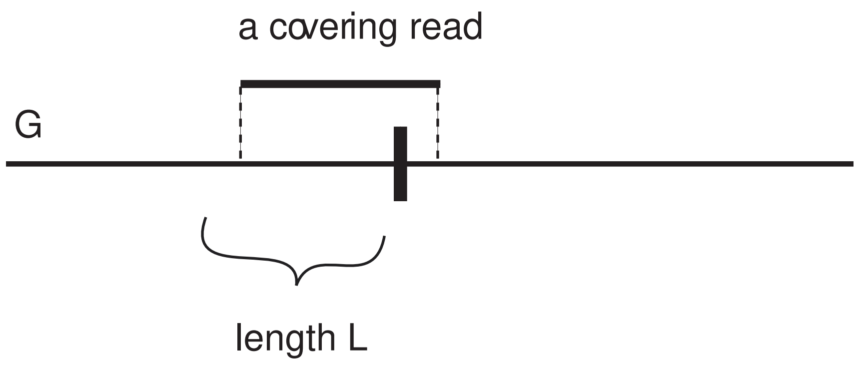

Figure 5.3: When a read 'covers' a given base (at center).

What are the chances that a read starts at a given base? Start point is uniformly distributed at random, hence chance is $1/G$. Chance of a sampling *not* starting at a given base is then $1 - 1/G$. For R reads, the chance of no read starting at a base is $(1 - 1/g)^R \approx 1 - R/G$ for $R \ll G$.

Now, the probability that no read covers a given base is the probability of no base starting at that base, or any base within L bases upstream of the given base in question.

Reads starting at different bases are independent events, hence the probability of no reads starting at a given L bases is simply,

$$(1 - \frac{R}{G})^L = (1 + (-\frac{R}{G}))^{-c(-G/R)} \tag{5.2}$$

For $R/G \to 0$, or $u \to \infty$ with $u = G/R$,

$$\lim_{n \to \infty} (1 + \frac{1}{u})^{cu} = e^{-c} \tag{5.3}$$

Hence, the probability of a base having *not* been sampled goes as e^{-c}, and the percent of total coverage goes as $1 - e^{-c}$.

Question 2: How many gaps are there?

We consider a 'gap' to be a break in the coverage of reads. However, functionally, we have coverage only when we can confidently match two consecutive reads together, which requires more than a one base-pair overlap between reads.

Hence, anything less than a θL base overlap we consider a gap, where θ is some fraction. For the sake of throwing out numbers, assume we want an overlap of fifty bases (the 4^{-50} chances of such a matching resulting randomly are pretty slim), meaning that for reads of length $L = 500$, we need about a 10% overlap ($\theta = .1$).

So, given one read along the sequence, what are the chances that the subsequent read does not start soon enough such that our overlap threshold is met? From our previous question, we know that the probability a read does not start at a given base-pair is $1 - R/G$. For an overlap of θL to happen, the read must start at any of $(1 - \theta)L$ positions. So, for a read *not* to start at any of those positions we have a probability of $(1 - R/G)^{(1-\theta)L}$. Given R reads, then the number of non-reads along the critical region is,

$$(R - 1)(1 - R/G)^{L(1-\theta)} \approx (R - 1)e^{-c(1-\theta)} \approx Re^{c(1-\theta)} \tag{5.4}$$

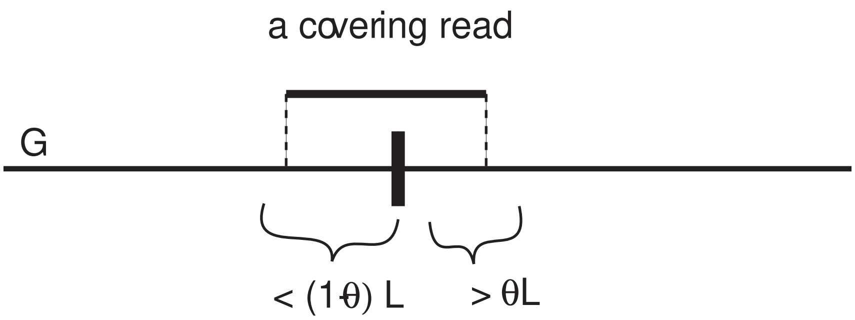

Figure 5.4: Dynamics of the θ gap.

Note, we made the same approximation to an exponential as above. Also, we first multiplied by $R-1$, which is the number of reads followed by a read(which is $\gg 1$ anyhow). This we write in terms of the parameter we can tweak c, $G/Lce^{-c(1-\theta)}$. The resulting function has a maximum at $1/(1-\theta)$.

Question 3: How long is a gap, on average?

For a gap of length k, we need to have had no reads start for k positions, followed by a read at the $k+1$th position. The probability of such an event is $(1-\frac{R}{G})^k\frac{R}{G}$. The average length is then this probability times length, summed over all possible lengths,

$$\sum_{k=1}^{G} k(1-\frac{R}{G})^k\frac{R}{G} = \frac{R}{G}\sum_{i=1}^{G} ka^k = \frac{a}{(1-a)^2}\frac{R}{G} = \frac{L}{c} - 1 \approx \frac{L}{c} \tag{5.5}$$

having set $a = 1 - R/G$ and using the definition of c. So the average length of a zero-gap goes inversely as the coverage.

Question 4: How big is a zero-contig, on average?

From above we know that the number of gaps, and hence roughly of contigs, is Re^{-c}. Likewise, from Q3 we know each gap has an average length of L/c. So the amount of the sequence not consumed in gaps (and hence forming part of a contig) is $G - \frac{L}{c}Re^{-c}$. Dividing by the number of contigs gives us an average

$$\frac{G - \frac{L}{c}Re^{-c}}{Re^{-c}} = \frac{G}{R}e^c - \frac{L}{c} \tag{5.6}$$

We see that the average contig size grows exponentially in the coverage.

5.4 Algorithmic approaches

Make an appeal to parsimony in sequencing (Occam's razor): give me a string, as short as possible, that explains the data. More formally, find the shortest string C such that $\forall$ reads v_i, $\exists$ (b_i, e_i) such that $C[b_i, e_i] = v_i$.

Full coverage is assumed in the problem statement, as, if there were any gaps, they would be removed to satisfy the minimum length condition.

This is the shortest common super-string (SCS) problem, which is NP-hard. Approximation algorithms get to with $1.5x$ of optimal, but that is unsatisfactory for biological work.

Abandon perfectionism, and instead demand an approximate matching[M95]. In problem statement above, replace equality with $C[b_i, e_i] \tilde{\epsilon}$ v$_i$. The relation $\tilde{\epsilon}$ represents an approximate matching wherein the mismatch is beneath a threshold ϵ(e.g. no more than $(e_i - b_i)\epsilon$ bases not matching).

All is not well with that, however. If our original sequence is the below,

Figure 5.5: The true sequence of interest. X represents a repeat.

Then the resulting matching is the following,

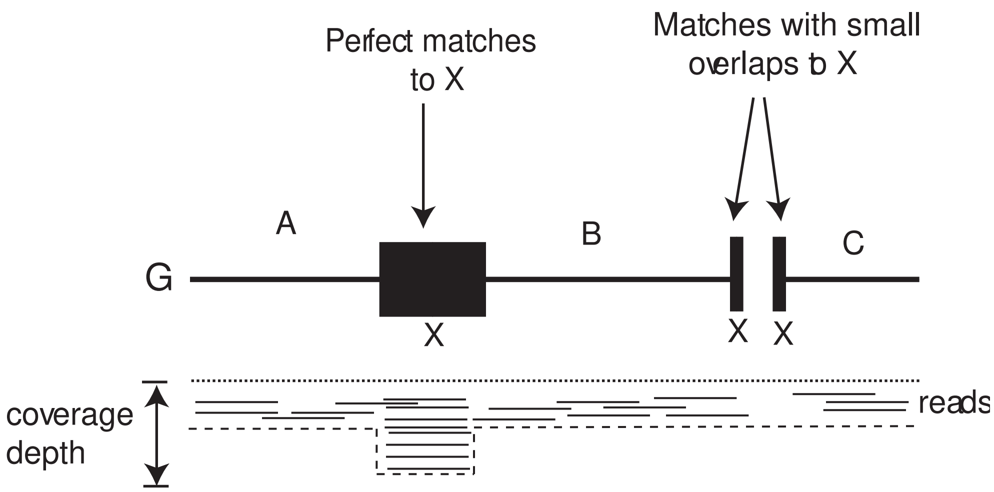

Figure 5.6: Matching under shortest sequence constraint. The over matching under the left X corresponds to perfectly matched segments, while the nubs to the right reflect segments with some overlap to X. The push to shorten the sequence results in a significant misalignment.

As is clear[screwy], the alignment is far from ideal. Rather than demanding a minimum sequence, we try to find most probable alignment based on the random fragmentation (and hence random coverage) of the sequence fragments. The probability density function for the event of a reading starting at a position is essentially $1/G$, and hence the cumulative density function of events occuring between the sequence start and some other position x_i hence increases linearly with length along G. In other words, the further along G that we go, the more segment starts we expect to see, and the rate of discovery of starts should be constant as we assume a random and unbiased fragmentation process. If our resulting alignment does not conserve that property (and we see bunching or gaps of segment starts), then we suspect our alignment is not valid. This is not easy to put on an algorithmic footing, and has not been done to date.

We quantify such a difference via the Kolmogorov-Smirnov test statistic which measures the deviation between an observed sample and an idealized source distribution (the random distribution in our case). The intriguing thing is that the K-S statistic is distribution free, and hence assumes nothing about the underlying source distribution.

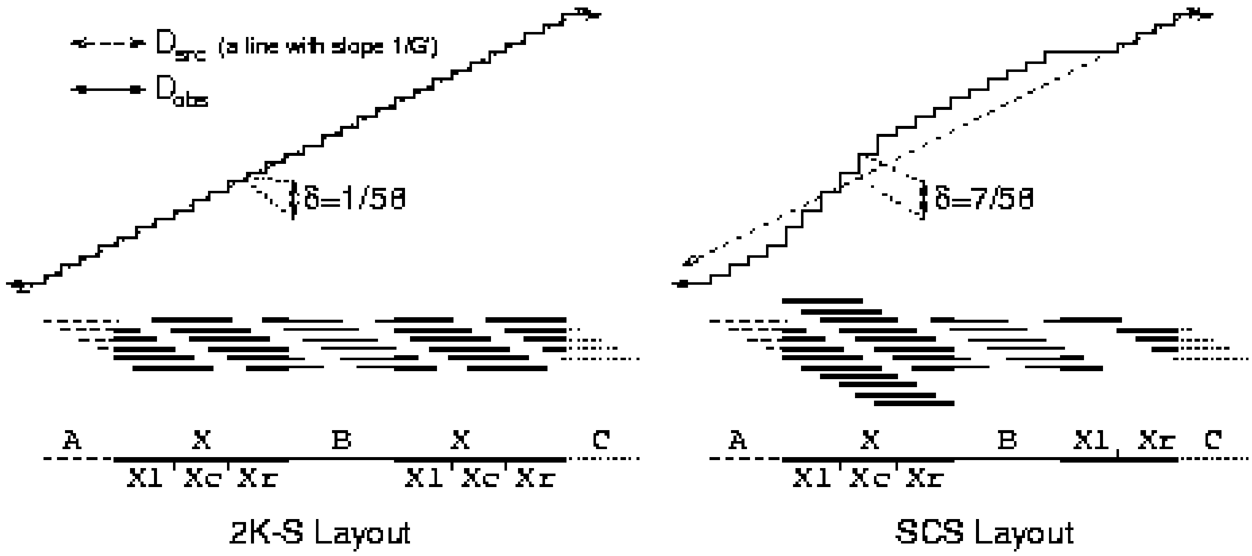

Figure 5.7: Both plots show coverage and segments starts along a $G = 58$: position on x-axis and number of starts along y. The left is the ideal, with starts occuring with rate 1/58 along G. On the right, alignment has skewed the segments starts such that a bunching results in the plot overshooting the theoretical norm (with a slope of 7/58), and then undershooting later at a relative gap.

References

[M95] E.Myers, Toward Simplifying and Accurately Formulating Fragment Assemby , *Journal of Computational Biology* 2,2 (1997), pp.275-290.

[KM95] J. Kececioglu and E. Myers, Combinatorial Algorithms for DNA Sequence Assembly, *Algorithmica* 13, 1-2 (1995), pp.7-51.